Becoming reciprocal ?

Linzhuo Li

1. Research Question

Question and answering(Q&A) communities have been increasingly popular in cyber space. Members in these communities such as Quora, Stack Exchange, Google Answers contribute to build an ever expanding online knowledge repository regarding almost every aspect of life, and knowledge cumulated in these communities have greatly benefited millions of people online and offline.

Previous studies of knowledge sharing in Q&A communities often take a non-interactive perspective. They often fit the growth curve of questions/answers, or look at the general patterns of user activeness, user lifetime, etc. By doing this they were able to capture a lot of the characteristics of these communities. However, they tend to neglect an important and interesting aspect of questioning and answering as human interaction: how users choose the roles of questioner and answerer. For interactions to be sustaining, askers from the “demand" side with answerers from the “supply” side have to match. This “demand-supply” metaphor indicates that the question of role distribution is non-trivial and often can affect the sustainability of a knowledge community. Moreover, user activeness follows a long tail distribution and most questions and answers are produced by only a small proportion of active users repetitively engaging in Q&A interactions. These active users may only ask questions to acquire knowledge, or only provide answers, they may also both being a questioner and answerer. Their choice collectively determines the demand and supple of knowledge, but it is not well studied how their roles form and evolve over repetitive interactions.

Specifically, I am interested in whether users will become “reciprocal” overtime. By “reciprocal" I mean a state of both getting knowledge from other users and

contribute their own knowledge to others. It has been assumed that on Q&A websites, users operate based on “generalized reciprocity”(Wasko & Faraj, 2000): "help given to one person is reciprocated by someone else and not by the original recipient of the help" (Ekeh, 1974). And it has been argued that generalized reciprocity is important for community growth. However, no empirical studies are able to quantify and describe to what extend reciprocity exists and whether generalized reciprocity can emerge from interactions, or whether users are becoming reciprocal.

2. Big data

Therefore, my research question is whether Q&A interactions can generate/foster reciprocity among users. To answer this question, I plan to use the open source data dump from Stack Exchange. Stack exchange is a platform for 162 various knowledge communes. The observational data from Stack Exchange has an almost complete track of users behavior (including their personal information, their question and answering history, their votes, their comments and their reputations). This data is made public by the Stack Exchange to benefit researchers, and it has been used in quite a few studies. Currently the data dump has records of histories up to September 12th and it has a size of about 40GB.

The project can illustrate the good characteristics of big data. First, it is complete relative to other small data. It has a wide variety of knowledge communities that can compared or analyzed to look for a general pattern. It also has the information from the very beginning of each community, and one can learn users’ behaviors by the time they arrive at Stack Exchange. Second, issues such as data drifting, sampling bias can be measure and controlled because there are informations about users’ gender, location, age, career and so on. Third, the data is non-reactive, clearly. Fourth, compared with other big data, the archival of Stack Exchange keeps an elegant xml format with almost every features will documented, and

according to previous users, it doesn't need a lot of data cleaning, in other words, it is not as "dirty" as other big data.

There is a bad aspect of the Stack Exchange data, namely, while most communities are relatively small (with 10^2-10^4) users, one community is extremely large: the stack overflow community has about 10^6 users. Therefore, it is not directly clear that interactions on stack overflow are comparable with other smaller communities. In terms of my question, it may be possible that fine-grained interactions can be better observed on stack overflow than in other communities, and because stack overflow has more users and more observations, patterns on stack overflow might be more stable than in other communities. So I need to be careful if I want to make generalizations using some results only from large communities.

3. Feasibility and other concerns

Because the data is open source and can be downloaded directly by everyone, it saved a lot of effort doing data collection.

Nevertheless, I noticed that the data contains personal information such as user's Facebook or twitter accounts, and therefore one may link their user behaviors with their real personal infos on Facebook/Twitter, which may cause concerns. I plan to not draw too much on that part of information, partly because I am not look particularly at their real personal attributes or how their real life role correlate with their Q&A behaviors, and for some personal information such as their gender, age, etc that I do need to look at, I will convince IRB that these information will not be used to do harm to users in my research or put them at risk, for example, I will not list users' nicknames in the research. There will not be any conclusions based on personal characteristics.

Ekeh, P. P. (1974). Social exchange theory: The two traditions. Harvard Univ Pr.

Wasko, M. M., & Faraj, S. (2000). “It is what one does”: why people participate and help others in electronic communities of practice. The Journal of Strategic Information Systems, 9(2), 155-173.